Willis MJ, vonStosch M. Inference of chemical reaction networks using mixed integer linear programming. *Computers and Chemical Engineering* 2016, 90, 31-43.

Inference of Chemical Reaction Networks using Mixed Integer Linear Programming

Mark J. Willis and Moritz von Stosch
School of Chemical Engineering and Advanced Materials
University of Newcastle
Newcastle upon Tyne, NE1 7RU.
mark.willis@ncl.ac.uk, moritz.von-stosch@newcastle.ac.uk

Abstract - **The manual determination of chemical reaction networks (CRN) and reaction rate equations is cumbersome and becomes workload prohibitive for large systems. In this paper, a framework is developed that allows an almost entirely automated recovery of sets of reactions comprising a CRN using experimental data. A global CRN structure is used describing all feasible chemical reactions between chemical species, i.e. a superstructure. Network search within this superstructure using mixed integer linear programming (MILP) is designed to promote sparse connectivity and can integrate known structural properties using linear constraints. The identification procedure is successfully demonstrated using simulated noisy data for linear CRNs comprising two to seven species (modelling networks that can comprise up to forty two reactions) and for batch operation of the nonlinear Van de Vusse reaction. A further case study using real experimental data from a biodiesel reaction is also provided.**

Keywords - chemical reaction network; structural optimisation; kinetic fitting; mixed-integer-linear programming

1.0 Introduction

A major barrier in the transition from chemistry research to process development is that more quantitative information, regarding a chemical synthesis, is needed. For process development the over-riding concern is to mathematically characterise the route from reactants to products, which often occurs over several reaction steps, with the involvement of measurable intermediates. The objective is to develop stoichiometric and kinetic descriptions of chemical reactions as opposed to obtaining a detailed mechanistic understanding of a synthetic route. For multiple reaction systems, this is referred to as a chemical reaction network (CRN). A mathematical model of a CRN, written as a set of coupled non-linear ordinary differential equations (ODEs) describing the dynamic behaviour of the system, instantiates a CRN in commercial process simulation and optimisation software. Software of this nature is required for numerous reasons including, accurate and economic plant design and process optimisation (Maria, 2004) and so methods, tools and procedures for rapidly establishing a CRN using experimental data are desirable.

In particular, methods that can be applied to data obtained from reaction systems operating far away from chemical or biochemical equilibrium are of interest. This is because batch and semi-batch reactors – rather than continuous stirred tank reactors (CSTRs) – tend to be used in the fine chemical and pharmaceutical

industries during the new chemical entity development lifecycle. Furthermore, the increased uptake of high throughput technologies e.g. automated robotic workstations for performing many experiments in parallel, coupled with improved sensor technology is likely to provide an increase in the quantity and quality of non-equilibrium experimental reaction data.

The work by Aris and Mah (1963) has been the basis for many stoichiometric and kinetic modelling studies aimed at CRN determination. One of the earliest advances being made by Bonvin and Rippin (1990) who proposed target factor analysis (TFA). This may be used to identify the number of linearly independent reactions. It may also be used to test suggested reaction stoichiometry is consistent with experimental data. Using TFA as a basis Brendel et al. (2006) and Bhatt et al. (2012) demonstrate an incremental identification strategy for CRNs. In their step-wise procedure, reaction stoichiometry (the CRN structure) is identified using TFA and then kinetic model identification strategies are used to determine the most appropriate ODE model (specifically, the structure and parameters of the rate laws). This decomposition of the problem allows for systematic development of a CRN. However, as stepwise methods are essentially local search operators they may produce sub-optimal solutions (local minima to the global optimisation problem).

A contrasting approach is to parameterise a suitable ODE model structure directly from the observed data and to use the resulting model to infer network properties. Domain dependent knowledge can be exploited to narrow the network search space, by restricting the form of the ODE model used to explain the dynamic behaviour of the CRN. For instance, (bio) chemical reactions, occurring in well mixed, relatively dilute, homogeneous phases – such as may be found in controlled laboratory batch and fed-batch experiments – typically obey the law of mass action kinetics. This allows a class of physically interpretable ODE models with pseudo-linear properties to be formulated, enabling classical regression techniques to be applied to the CRN model search process. While Crampin et al. (2004), Searson et al. (2007), Srividhya et al. (2007), Burnham et al. (2008), Searson et al. (2012), Hii et al. (2014) demonstrate the potential of this approach the identification methods used fail to consistently predict an underlying network structure. Potential explanations for this are that, **a)** Crampin et al. (2004), Srividhya et al. (2007), Burnham et al. (2008) use step-wise identification procedures which are susceptible to local minima, **b)** while Searson et al. (2007), Searson et al. (2012), Hii et al. (2014) use evolutionary algorithms (EAs) structural constraints are difficult to incorporate into any EA (it is generally left to the objective function to manage and quantify possible structural infeasibility rather than directly imposing these constraints as part of network search). Therefore, an EA will perform poorly if the search space is highly constrained.

MILP (and variants such as mixed integer quadratic programming, integer linear programming etc.) has been used for many years for process synthesis, scheduling and control and a vast amount of literature is available, e.g. Grossmann (1985), Achenie and Biegler (1990), Raman and Grossmann (1991, 1992), Floudas and Lin (2005), Moro and Grossmann (2013) are a selection of examples of the literature in this area. To synthesize process flowsheets using MILP, a superstructure e.g. see Achenie and Biegler (1990) is often used. This is constructed to contain all possible alternatives of a potential process flowsheet of which the optimal solution belongs. The use of a superstructure for process synthesis has proved an effective tool in many application studies. Adopting this approach for CRN elucidation, defines a global model structure (which can be represented as a digraph) consistent with all possible kinetic rate terms arising from

elementary chemical reactions. In effect this simplifies the difficult task of simultaneous structure and parameter estimation to one of just parameter estimation, where correct estimation of the parameters (links within the digraph) is, in principle, sufficient to deduce the structure of the underlying reaction network.

In this work, parameter estimation is achieved through minimization of the sum of the absolute errors (the L_1 norm) between measured and predicted species concentrations. In the absence of additional constraints on the structure of the system equations any identification strategy will normally over-fit the observed data with terms being included which model measurement noise rather than actual system dynamics. This would have negative effects on both the portability of the model (its ability to model different instances of the system) and the interpretability of the model (vital for network identification). Therefore, it is preferable to introduce additional information into the cost function in order to balance the trade-off between model complexity (in this case, the number of reactions) and how well the model fits the data. A number of regularisation techniques are available, including ridge regression (Hoerl and Kennard, 1970) and the Least Absolute Shrinkage and Selection Operator (LASSO) proposed by Tibshirani (1996) e.g. see a review by Hesterberg et al. (2008). The method proposed in this paper is conceptually similar to these techniques however, a set of binary variables (associated with each of the parameters of the model) are used to perform regularisation rather than the parameters themselves. The binary variables provide a normalised entropy measure (independent of the magnitude of the regression parameters) and are directly related to the number of chemical reactions within the network. Whilst determination of these binary variables increases the number of parameters being identified as part of the optimisation process, they are also used to indicate 'options' (e.g. between different reactions or species combinations) effectively turning 'on' or 'off' alternative solutions to an optimisation problem. This provides a flexible identification framework that may be used to incorporate known information about the chemical species and reactions in the form of linear equality and inequality constraints. Furthermore, the MILP can be solved using fast, efficient and readily available commercial (and open source) software.

2.0 Chemical reaction networks

A CRN may be represented by a directed graph (or digraph). For example, Fig. 1 shows a weighted di-graph representation of the following CRN comprising four reactive species x_1, …., x_4 involved in three reactions (known as the Van de Vusse reaction scheme),

$$2x_1 \rightarrow x_2 \qquad (1)$$
$$x_1 \rightarrow x_3 \rightarrow x_4$$

The graph is labelled with nodes, $\mathbf{m} = [c_1 \quad c_1^2 \quad c_2 \quad c_3 \quad c_4]^T$ and there are connections between the nodes which are referred to as **edges** (or **links**) in the graph. The links have a specified direction - as indicated by the arrows. The nodes are associated with the concentration of chemical species and are defined using of the law of mass action kinetics[1], assuming in principle, that each of the reactions are reversible. A constant of proportionality (the isothermal rate constant for the particular reaction) is the label on the edge.

[1] Using mass action kinetics to define the monomials (and hence the nodes of the weighted digraph) is not a restriction of this technique. If explicit knowledge of alternative rate law structures exists then they may be used to define the node(s).

The topology of a digraph may be represented using a matrix and there are a number of such matrices that can be associated with any graph, e.g. the adjacency matrix, the incidence matrix and the Laplacian matrix, see e.g. Agarwal and Singh (2009). In this work the Laplacian matrix which was first introduced by Kirchhoff (1847) in his article about electrical networks is used as it provides a compact representation of the dynamics of a CRN.

The weighted Laplacian matrix $\mathbf{K} = [k_{pq}]$ is a (n x n) matrix representing the topology of the network (the interactions between the n nodes within the digraph). The elements within **K** are defined as (where d_p is the sum of the values of the off diagonal terms within each row),

$$k_{pq} = \begin{cases} k_{pq} & if\ node\ p\ and\ q\ are\ incident \\ 0 & if\ node\ p\ and\ q\ are\ not\ incident \\ -d_p & if\ p = q \end{cases}$$

As the label on the edge is the isothermal rate constant, the rate of production (or consumption) of each of the chemical species **r** (N_S x 1) as a result of each of the chemical reactions may be defined as,

$$\mathbf{r} = \mathbf{K}^{\mathrm{T}}\mathbf{m} \tag{2}$$

If the chemical reactions take place in a homogeneous well mixed batch reactor the rate of change of these monomials $\dot{\mathbf{m}} = [\dot{m}_1 \quad , \dots , \quad \dot{m}_{N_m}]^{\mathrm{T}}$ with respect to time is,

$$\dot{\mathbf{m}} = \mathbf{r} = \mathbf{K}^{\mathrm{T}}\mathbf{m} \tag{3}$$

The derivative of the species concentrations $\dot{\mathbf{c}} = [\dot{c}_1 \quad , \dots , \quad \dot{c}_{N_S}]^{\mathrm{T}}$ may then be defined as,

$$\dot{\mathbf{c}} = \mathbf{N}^{\mathrm{T}}\dot{\mathbf{m}} = (\mathbf{KN})^{\mathrm{T}}\mathbf{m} \tag{4}$$

where **N** (N_m x N_S) comprises the stoichiometric amounts of each of the monomials that appear in the nodes of the weighted digraph. For the Van de Vusse reactions this gives,

$$\begin{bmatrix} \dot{c}_1 \\ \dot{c}_2 \\ \dot{c}_3 \\ \dot{c}_4 \end{bmatrix} = \left(\begin{bmatrix} -k_{14} & 0 & 0 & k_{14} & 0 \\ 0 & -k_{23} & k_{23} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -k_{45} & k_{45} \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \right)^{\mathrm{T}} \begin{bmatrix} c_1 \\ c_1^2 \\ c_2 \\ c_3 \\ c_4 \end{bmatrix} \tag{5}$$

There are three distinct aspects to (4) that contribute to the rate of change of each species concentration, **a)** the stoichiometric amounts of the monomials, **N**, **b)** the underlying topology of the CRN, which is contained within the weighted Laplacian matrix and **c)** the reaction rate functions embedded in nodes of the weighted digraph. The dynamic model (4) reveals an underlying linear character (3) within the system equations which is advantageous for systems analysis as the properties of linear ODEs are well known. This relationship between the graphical structure of a CRN and the dynamic description of a set of chemical reactions forms the basis of Chemical Reaction Network Theory (CRNT), initiated by Horn and Jackson (1972). CRNT has received attention in recent years as it provides mathematical tools to elucidate the qualitative dynamics, such as the stability and uniqueness of the CRN through analysis of the CRN model structure alone, i.e. independent of the values of the rate constants and the exact form of the rate

expressions, e.g. see Craciun and Feinberg (2005, 2006, 2010), Mincheva and Roussel (2007), Mincheva (2011).

2.1 Structural identification of chemical reaction networks

If the chemical reactions (and hence their rate law structure) are unknown, then any feasible monomial may be generated and used in system identification studies. Pragmatically however, assumptions may be made to generate a finite number of terms. For example, if it is assumed that **a)** the reactions are at most the result of bimolecular collisions and the total reaction order is no greater than two, **b)** the monomials are of integer order with respect to each species concentration, **c)** there are no terms consisting of purely zero order elements. Then, given N_S species, the N_M possible monomials describing all the feasible reactions between the species are,

$$\begin{matrix} c_j & (j = 1, \ldots., N_s) \\ c_i c_k & (i = 1, \ldots., N_s), (k = 1, \ldots., N_s) \end{matrix} \quad (6)$$

The monomials (6) may be used to construct an extended vector **m** (N_M x 1) representing all feasible mono-molecular and bi-molecular chemical reactions between the N_S species - for reactions (1) this gives[2],

$$\mathbf{m} = [c_1 \quad c_1^2 \quad c_2 \quad c_3 \quad c_4 \quad c_1c_2 \quad c_1c_3 \quad c_1c_4 \quad c_2^2 \quad c_2c_3 \quad c_2c_4 \quad c_3^2 \quad c_3c_4 \quad c_4^2]^{\mathrm{T}} \quad (7)$$

This defines a (N_M x N_M) weighted Laplacian matrix and a matrix **N** (N_M x N_S) comprising the stoichiometric amounts of the N_M monomials. In other words, it encodes within the dynamic equations (4) a CRN superstructure – see Fig. 2 – representing all possible combinations of reactions that could occur given the monomials (7). This is a highly redundant network where, each potential reaction that could transform reactants into products, possibly over several reaction steps, are included and interconnected with the rest. However, an ideal parameter estimation procedure applied to (4) using experimental data should give the non-zero values within **K**, i.e. discover the structure of the underlying CRN as well as values of the isothermal rate constants associated with each of the reactions.

As the weighted Laplacian (and hence the ODEs) are an over parameterized model, in order to develop a robust identification framework a two stage procedure is adopted in this work, **a)** a strategy is developed for reducing the number of monomials (nodes in the digraph superstructure) thereby reducing the size of the system identification problem, **b)** additional model constraints are introduced that allow the efficient identification of sparse representations of (4) using MILP.

3.0 Elimination of infeasible reactions from the search space

As the number of chemical species N_S increases, the monomials within **m** increases quadratically as,

$$N_M = 2N_S + \frac{N_S(N_S-1)}{2}, N_S > 2 \quad (8)$$

[2] For clarity, the vector has been ordered to show the monomials associated with (1) first, followed by the additional terms generated using (6).

A model reduction strategy is therefore required, in most cases, to eliminate redundant monomials and identify infeasible structural interactions. This may be achieved if the elemental make-up of the species is known identifying combinations of monomials that can produce mass balanced reactions. If **A** (N_S x N_A) is the atomic (or elemental) matrix, whose N_S rows represent the species and the N_A columns the distinct atoms (or molecular groups) that define the chemical species,

$$\mathbf{A} = \begin{pmatrix} a_{11} & \cdots & a_{1,N_A} \\ \vdots & \ddots & \vdots \\ a_{N_S,1} & \cdots & a_{N_S,N_A} \end{pmatrix}$$

The elemental balance constraints (9) may be defined,

$$\mathbf{LNA} = \mathbf{0} \tag{9}$$

Where, $\mathbf{L} = [l_{pq}]$ is a (N_M x N_M) Laplacian matrix whose elements are defined as (and again d_p is the sum of the values of the off diagonal terms within each row),

$$l_{pq} = \begin{cases} 1 & if\ node\ p\ to\ q\ \ produce\ a\ feasible\ (mass\ balanced)\ reaction \\ 0 & if\ node\ p\ to\ q\ do\ not\ produce\ a\ feasible\ (mass\ balanced)\ reaction \\ -d_p & if\ p = q \end{cases}$$

The off-diagonal elements of **L**, may be obtained by maximizing the objective function (10) subject to the constraints (9) and $l_{pq(q\neq p)} \in \{0, 1\}$.

$$max \sum_{p=1}^{N_M} \sum_{q=1(q\neq p)}^{N_M} l_{pq} \tag{10}$$

If any entire row of zeros is obtained within **L** then the monomial that the row represents may be removed from the feasible set of monomials as it does not produce a single (balanced) chemical reaction. For example, for the reactions (1) the atomic matrix is (in this case a vector of relative atomic masses),

$$\mathbf{A} = [1 \quad 2 \quad 1 \quad 1]^{\mathrm{T}}$$

Multiplication with the matrix **N** that contains the stoichiometric amounts of the monomials given in (7) yields,

$$\mathbf{NA} = [1 \quad 2 \quad 1 \quad 1 \quad 1 \quad 3 \quad 2 \quad 2 \quad 4 \quad 3 \quad 3 \quad 2 \quad 2 \quad 2]^{\mathrm{T}}$$

With respect to (7) it is immediately obvious by inspection that the ninth element within **NA** (having a value of four) cannot produce a balanced chemical reaction. Further inspection reveals that the sixth, tenth and eleventh elements may also be removed as the only feasible (balanced) reactions that they produce involve species x_2 as both a product and a reactant in the same chemical reaction. Therefore, the vector of monomials (7) is reduced to,

$$\mathbf{m} = [c_1 \quad c_1^2 \quad c_2 \quad c_3 \quad c_4 \quad c_1c_3 \quad c_1c_4 \quad c_3^2 \quad c_3c_4 \quad c_4^2]^{\mathrm{T}}$$

Four terms have been removed from the original vector generated using (6). Each term that has been removed corresponds to a reaction involving species x_2 (whose relative molecular mass is twice that of the others) where a balanced chemical reaction is not obtained. For chemical reactions (1) the remaining terms

in **L** and the corresponding digraph are shown in Fig. 3. Each zero represents an infeasible reaction given the elemental balance constraint. In terms of parameter identification, the unknown terms within **K** correspond to the non-zero elements in **L** - and each zero within **L** defines a zero within **K**. It may be observed for the reactions (1) that the degree of each node within this feasible (mass balanced) superstructure is typically small indicating that each monomial can only be paired with a limited number of the other monomials to define a feasible reaction.

4.0 Model parameter estimation

To avoid the numerical integration of (4) it is possible – prior to the network search procedure - to estimate the time derivatives for each of the N_S species from the measured concentration data. This can be achieved by fitting non-linear smoothing functions to the concentration measurements and analytically obtaining the derivatives at the required sample times from the fitted function (methods to do this are discussed in section 6), giving a vector of estimates of the time derivatives of the species concentrations at each time, t = 1, 2, …, N_t, $\mathbf{S}_t = (S_{c_1,t}, \dots, S_{c_{N_S},t})^{\mathrm{T}}$. The system of ODEs (4) are consequently reduced to a set of algebraic equations which may be solved for the unknown parameters within the weighted Laplacian, by means of multiple linear regression (MLR) rather than computationally expensive methods that require numerical solution of the set of ODEs.

As opposed to using a least squares objective function (minimizing the squared error between an output and a predicted output) in this work the objective function used is the sum of the absolute errors (the L_1 norm),

$$J_{LAD} = \sum_{t=1}^{N_t} \|\mathbf{S}_t - (\mathbf{KN})^{\mathbf{T}}\mathbf{m}\|_1 \tag{11}$$

Also known as Least Absolute Deviations (LAD), (11) finds application in a number of system identification studies as it is robust (resilient to outliers in the data) as it gives equal emphasis to all observations. Moreover, it may be formulated as a linear objective function. Rewriting (11) using a vector of artificial variables, $\boldsymbol{z}_t = (z_{1,t}, \dots, z_{N_s,t})^{\mathrm{T}}$ gives (12) which is a linear cost function – and if this is minimised - subject to the constraints (13) is equivalent to (11).

$$J_{LAD} = \sum_{t=1}^{N_t} \sum_{j=1}^{N_S} z_{j,t} \tag{12}$$

$$\begin{aligned} \boldsymbol{z}_t &\geq \mathbf{S}_t - (\mathbf{KN})^{\mathbf{T}}\mathbf{m} \\ \boldsymbol{z}_t &\geq (\mathbf{KN})^{\mathbf{T}}\mathbf{m} - \boldsymbol{S}_t \\ \boldsymbol{z}_t &\geq \mathbf{0} \end{aligned} \tag{13}$$

The optimum CRN structure will generally be comprised of significantly less edges than the full digraph with the same node set, i.e. **K** will be a sparse matrix. Any optimisation problem should therefore include the ability to select a subset of the reactions (nodes) and their interconnections within the original superstructure. It is possible to directly minimise the number of non-zero terms within **K** by introduction of a set of binary variables $l_{pq(p\neq q)} \in \{0,1\}$ to each element $k_{pq\,(p\neq q)}$ in the weighted Laplacian matrix. The binary variables take a value of one *if and only if* the corresponding weight within **K** (the isothermal rate constant) is greater than zero,

$$l_{pq} = 1 \;\forall\; k_{pq\,(p\neq q)} > 0, (p = 1, \dots, N_M), (q = 1, \dots, N_M) \tag{14}$$

To specify the values of l_{pq}, (which define the Laplacian matrix, **L**) the logical statement (14) must be transformed into linear inequalities to maintain the integrity of the MILP. It is straightforward to verify e.g. see Hadjiconstantinou and Mitra (1994), that this is achieved using the linear constraints,

$$\left.\begin{matrix} k_{pq} \geq \varepsilon l_{pq} \\ k_{pq} \leq M l_{pq} \end{matrix}\right\} \quad (p = 1, \dots, N_M), (q = 1, \dots, N_M) \tag{15}$$

where $0 < \varepsilon \ll 1$ (in other words estimated rate constants below ε are taken as zero) and M is the maximum limit specified for the isothermal rate constants. Therefore, the minimization of the number of non-zero components within **K** amounts to the minimization of the sum of the corresponding binary variables, i.e. the following cost function may be defined,

$$J = \sum_{t=1}^{N_t} \sum_{j=1}^{N_S} z_{j,t} + \lambda \sum_{p=1}^{N_M} \sum_{q=1}^{N_M} l_{pq} \tag{16}$$

Where λ is a weighting parameter (known as the model regularisation parameter) used to balance the trade-off between model complexity (in this case, the number of reactions as, $N_R = \sum_{p=1}^{N_M} \sum_{q=1}^{N_M} l_{pq}$) and how well the model approximates the experimental data. This is referred to as L_0 – norm regularisation as the binary variables determine the cardinality of **K** (i.e. a count of the number of non-zero values in **K**). The cost function (16) is similar to (17) which somewhat corresponds to the LASSO regularisation approach.

$$J = \sum_{t=1}^{N_t} \sum_{j=1}^{N_S} z_{j,t} + \lambda \sum_{p=1}^{N_M} \sum_{q=1}^{N_M} k_{pq} \tag{17}$$

The distinct difference being that (16) uses a penalty term that has binary elements (corresponding to the number of reactions) while (17) has a penalty term that corresponds to the sum of the model parameters (the isothermal rate constants). As the isothermal rate constants will always be positive, the penalty in (17) is the L_1 – norm of the estimated model parameters. The use of this penalty term for model parameter (structure) selection is therefore referred to as L_1 – norm regularization.

5.0 Model structure selection

The choice of the optimal regularisation parameter (λ) is an important issue and the normal procedure adopted is to find process models that correspond to a range of the regularisation parameters (referred to as a solution or regularization path). The preferred model structure is then one that optimises a criteria such as the Akaike (1974) information criterion (AIC), or the Bayesian information criterion (BIC) or Schwarz criterion (Schwarz, 1978). Both these criteria trade-off model complexity against how well the model approximates the data. It is known that AIC-based methods are not consistent for model selection as irrelevant model parameters tend to be selected e.g. see Shao (1993). However, the BIC has been shown to be successful, e.g. see Wang, Li, and Tsai (2007); Zou, Hastie, and Tibshirani (2007). This criterion may be described by,

$$BIC = -2\, ln(lik) + ln(n) df$$

Where $'lik'$ is the maximum value of the likelihood function of the model, df is the number of parameters (degree of freedom) of the model and n is the number of data points. Following (Gao, 2008) using the LAD cost function (11), the BIC (cost function) is (where J_{LAD} is the LAD cost function defined earlier),

$$BIC(\lambda) = n\ln(J_{LAD}/n) + ln(n)df \quad (18)$$

An optimal model structure corresponds to the regularization parameter λ that minimizes (18). Therefore, either cost function (16) or (17) may be minimised for a range of λ values in order to determine the regularization path (or landscape) and the corresponding values of (18) are calculated to determine the optimal CRN structure. For LASSO, Zou, Hastie, and Tibshirani (2007) prove that the number of non-zero coefficients within the model is an unbiased estimator of the model degree of freedom, df. For our MILP strategy this is the sum of the binary variables associated with each model parameter; for the LASSO implementation (17) the number of non-zero values within the parameter vector may be heuristically determined and summed.

6.0 Implementation aspects

The flowchart, Fig. 4 demonstrates our algorithm workflow. The first step, assuming data are in an appropriate format, is the smoothing of each of the species concentrations. The smoothed data (either obtained automatically or by data 'eye-balling' and iteration) are then used as a replacement for the raw (noisy) data. Each smoothed signal is then used to calculate the derivatives of the respective species concentrations. In parallel, the weighted Laplacian is defined through specification of the monomials of the reaction rates terms; known system information is taken into consideration at this stage (i.e. if the elemental matrix is known then a subset of mass balanced connections can be specified), if it is not known the model can be taken as the fully connected graph. The MILP is solved (in this work we have used the function 'intlinprog' in MATLAB) subject to the constraints (9), (13), (15). The resulting parameter estimations (and hence network structures) are analysed across the regularization landscape. The premise is that through the use of (18) a unique CRN structure will be identified. However, if the situation were to arise that there are a number of solutions with similar BIC criterion results further data and / or additional system information would be needed (typical of any model development process).

6.1 Estimation of derivatives

Obtaining a good approximation of the first derivative of the times series concentration profile for each chemical species is crucial to the success of the method. This estimation, referred to as the inverse problem, is generally ill-posed, i.e. small errors in the concentration data can be amplified to large errors in the derivatives. The normal strategy adopted is to approximate the functional relationship between the species concentration and time, $c_i = f(t)$ over the time course of an experiment. Determining an accurate (smooth) estimate of the species concentrations then allows extraction of the appropriate derivatives. There are a number of parametric model forms that may be used to approximate the concentration data, including rational polynomials, smoothing splines, artificial neural networks etc., see e.g. Hosten (1979), Mata-Perez and Perez-Benito (1987), Kamenski and Dimitrov (1993), Voit and Almeida (2004), Bardow and Marquardt 2004), Marquardt (2005). In this work we use the curve fitting toolbox within MATLAB to fit a number of model forms (including rational polynomials and smoothing splines). The approach being to check by 'eye-balling' the data fit that the technique provides a good estimate of the respective species concentrations. First derivative data from the fitted curves is then calculated through differentiation of the fitted curves with respect to time.

7.0 Case Studies

In this section of the paper, three case studies are presented. The case studies have been designed to highlight the CRN elucidation procedure under a range of known experimental conditions before applying the technique to an actual experimental investigation. The first case study is used to demonstrate the estimation of the parameters within the weighted Laplacian in the absence of the derivative estimation highlighting the performance of the method under conditions of increasing CRN size and measurement noise. The second case study is an application to a simulated Van de Vusse reaction. This incorporates data smoothing and derivative estimation in order to demonstrate the effect of computing the derivatives numerically. The consequence of errors in the derivative estimations is evaluated. Finally, we apply our technique to real experimental data - a transesterification reaction - in order to highlight practical aspects of the CRN identification process.

7.1 Monte Carlo simulation of (linear) CRN's

The simulations are randomly generated (linear) CRN's (i.e. no nonlinear monomial terms) where the dimension of the Laplacian is randomly selected to be a (N_mxN_m) matrix with N_m set between 2 - 7 (corresponding to numbers of reactions of between 2 and 42). It is assumed that the general model (4) may be represented by,

$$\boldsymbol{y} = \mathbf{K}^{\mathrm{T}}\boldsymbol{u}$$

For each simulation, the sparsity of the CRN (number of zero parameters within **K**) was set randomly to a figure between 50 – 80%. Randomly generated input data $\boldsymbol{u}$ (n = 100) of zero mean and standard deviation of one was used to generate the output data $\boldsymbol{y}$ using the known CRN structure. White noise was then added to each of the model outputs. The objective, given the input data and the noisy model outputs was to estimate the parameters within **K.**

The results of one hundred Monte Carlo simulations (for each white noise level considered) are presented in Table 1. MAE is the mean absolute error between the actual outputs and the outputs of the CRN generated by the MILP over all the Monte Carlo simulations (the MAE is used as the cost function for the MILP is defined in terms of absolute error), $\left\|\hat{b} - b\right\|_2^2$ is the 2- Norm between the identified model parameters and the true model parameters used to generate the data. A false positive model is defined as a model that includes at least one additional parameter when compared to the true model. A false negative model is defined as a model that has at least one missing parameter when compared to the true model. The table summarizes the performance of the MILP using the optimal regularization parameter as defined by the BIC criterion (18).

At noise levels that would typically be expected of measured chemical species (< 10%), the BIC defined weighting correctly discovers the true CRN in 100% of the simulations (noise levels of 5 and 10%). With 20% white noise the BIC criterion enables 92% of the correct model structures to be identified (with 8% false positive models). At a noise level of 30%, 73% of the structures are identified correctly with both false positive and false negative models being identified; with some of the models being both false positive and false negative. The values for the two norm of the difference between the true and estimate model

parameters $\left\|\hat{b}-b\right\|_2^2$ increase with an increasing noise level due to the falsely identified reactions (with parameter values being different to their true value of zero) as well as due to the effect on the estimation of the model parameters as a consequence of the noise.

	MAE	$\left\|\hat{\boldsymbol{b}}-\boldsymbol{b}\right\|_2^2$	**False Positive (%)**	**False Negative (%)**	**True Model Structure (%)**
~5% Noise	1.086 ($\pm$ 0.715)	0.020 ($\pm$ 0.034)	0	0	100
~10% Noise	2.183 ($\pm$ 1.579)	0.077 ($\pm$ 0.11)	0	0	100
~20% Noise	5.163 ($\pm$ 3.31)	0.3816 ($\pm$ 0.51)	8	0	92
~30% Noise	6.79 ($\pm$ 4.72)	0.912 ($\pm$ 1.35)	22	9	73

Table 1. Performance L_0 – norm regularization using an optimal regularization weighting defined via the BIC over one hundred randomly generated simulations (the values in parenthesis are the standard deviations).

To demonstrate the nature of the regularization path for the L_0-norm approach to model regularization, the following CRN was simulated (species x_1 reacts reversibly to three product species x_2, x_3 and x_4),

$$\mathbf{K}=\begin{bmatrix}-3 & 1 & 1 & 1\\ 2 & -2 & 0 & 0\\ 3 & 0 & -3 & 0\\ 3 & 0 & 0 & -3\end{bmatrix} \tag{19}$$

The input data to the CRN comprised one hundred samples of randomly generated values taken from a normal distribution with a mean of zero and standard deviation of one. This data was used to generate the model output data using the known CRN structure. 10% white noise was then added to the model outputs. Minimizing (16) subject to the constraints (13) and (15) and constraints on the limits of the model constants, $0 \le k_{pq} \le 3.5$, the regularization parameter was varied between $\lambda = 0.0$ and 100 in increments of 0.1. The identified parameter values are shown in Fig. 5 (where the numbers 1 – 12 indicate parameter values ordered down successive rows of the Laplacian excluding the diagonal). It may be observed that at small values of λ some of the parameters within the CRN that should be zero have small non-zero values. However, once a sufficiently large value of λ is applied, the correct model structure is estimated (the optimal BIC have a weighting of $0.15 \le \lambda \le 88$) giving the estimated weighted Laplacian,

$$\hat{\mathbf{K}}=\begin{bmatrix}-2.97 & 0.98 & 0.98 & 1.01\\ 1.93 & -1.93 & 0 & 0\\ 3.02 & 0 & -3.02 & 0\\ 3.05 & 0 & 0 & -3.05\end{bmatrix}$$

The parameters demonstrate a slight bias from the true values – a consequence of the noise and limited number of data points, not the use of the regularization weight (increasing values of λ does not affect the parameter estimate until there is a change in the model structure). For this particular simulation it takes large values, $\lambda \ge 88$ before an additional term is dropped out of the model, causing an increase in prediction error and a subsequent rise in BIC criterion.

7.2 A simulated Van de Vusse reaction scheme

Suppose that experimental studies have been performed in a batch reactor aimed at the discovery of the (unknown) structure of the chemical reactions (1). The profiles of the concentration versus time of the measured chemical species are shown in Fig. 6. This data has been generated from the ODE description of the system (5) with the isothermal rate constants, k_{14} = 0.5 min^{-1}, k_{23} = 1.0 dm^3.kmol^{-1}.min^{-1} and k_{45} = 2.0 min^{-1}. Initially pure x_1 with a concentration of 2 kmol.dm^{-3} exists in the reactor. To simulate the effect of measurement error, the sampled concentration values for each species were independently corrupted with additive 10% white noise.

The rational polynomial described by (20) was found to fit the species concentrations well (other options including a smoothing spline were tried).

$$\hat{c}_{j,t} = \frac{a_1+a_2t+a_3t^2}{b_1+b_2t+b_3t^2} \ (j = 1, \dots N_S) \qquad (20)$$

In (20) $\hat{c}_{j,t}$ is the estimate of the j^{th} species measurement at time, t. The a_i $(i = 1, \dots, 3)$ and the b_i $(i = 1, \dots, 3)$ are model parameters that can be computed by a suitable nonlinear optimisation algorithm. The Levenberg – Marquardt nonlinear optimisation routine in MATLAB was used in this work. The optimisation process was repeated ten times with different initial starting points in an attempt to ensure the algorithm did not get trapped in local minima. The model with the best Sum of Squared Error (SSE) between the model estimate and the measured species concentration was retained. First derivative data from the fitted curves was then calculated through differentiation of (20) with respect to time. The estimate obtained for the derivative of the species concentration (and the actual value of the derivative taken from the known model equations) is shown in Fig. 7. Note that, although the actual derivatives are plotted alongside the estimates here for comparison, they will of course be unknown in practice.

It may be observed that for the first two species an accurate estimate of the model derivative has been obtained. For species three, the estimate is inaccurate at the initial points. The approximation of the derivative of species four is inaccurate for a number of points at the start of the batch. These inaccuracies were to be expected because (as discussed earlier) the estimation of the derivatives from measured concentration data is an ill-posed problem; this case study therefore allows an assessment to be made of the inaccuracies in derivative approximation on the performance of the proposed CRN elucidation strategy.

Fig. 8 shows the results obtained using the traditional L_1 –norm model regularization approach (the LASSO alike approach). The data used for CRN identification was specified as the smoothed species concentrations and the estimate of their respective derivatives. The cost function (17) was minimised subject to constraints (9) and (13) with the limits of the isothermal rate constants specified as, $0 \leq k_{pq} \leq 2.5$. The monomials used (and hence the Laplacian) corresponded to the reduced (mass balanced) reactions shown in Fig. 3. The regularization parameter was increased from an initial value of $\lambda = 0.01$ in steps of 0.01 to a value of 60. For each instance of regularization weighting (λ) the non-zero values of the parameters correspond to a possible CRN. It is obvious from Fig. 8 (and knowledge of the underlying model structure) that the correct CRN has not been identified. While the estimated value of k_{45} = 2.0 min^{-1} is initially close to being correct (at small λ values) there are still seven non-zero values of model parameters (corresponding to a CRN with

seven reactions). As the value of λ is increased the estimated value of k_{45} decreases and tends to zero before many of the other model parameters (reactions in the network).

Fig. 9 provides the results of the L_0 - norm regularization strategy. The data used for CRN identification was again specified as the smoothed species concentrations and the estimate of their respective derivatives. The cost function (16) was minimised subject to constraints (9), (13) and (15) with the limits of the isothermal rate constants specified as, $0 \leq k_{pq} \leq 2.5$. The regularization parameter was increased from an initial value of $\lambda = 0.01$ in steps of 0.01 to a value of 15. In Fig. 9 the BIC criterion is also shown for each value of the regularization parameter. The minimum of the BIC criterion occurs for all values of $\lambda \geq 0.4$. Increasing the value of λ beyond the final value shown would eventually reduce the number of model terms to zero at an increasing BIC cost. The minimum BIC corresponds to the CRN structure (5) with estimated values of the isothermal rate constants being, $\hat{k}_{14}$= 0.4964 min^{-1}, $\hat{k}_{23}$ = 0.984 dm^3.kmol^{-1}.min^{-1} and $\hat{k}_{45}$ = 1.985 min^{-1}.

7.2.1 Characterization of L_0 regularization of the Van de Vusse reaction scheme with varying noise levels

To fully characterize the performance of the MILP using the L_0- norm regularization approach (applied to the Van de Vusse reaction scheme) Monte Carlo simulations were again used. For a specified noise level (e.g. 10% white noise), one hundred randomly generated instances were added to the species measurements. The data were then smoothed and the derivatives estimated using (20). The smoothed values of the species and their derivatives were then used to estimate the optimal model structure and parameters using the MILP with constraints (9), (13) and (15) with the limits of the isothermal rate constants specified as, $0 \leq k_{pq} \leq 2.5$. The simulations were fully automated without the step of 'eye-balling' the smoothed fit to the species concentrations; the minimum of the BIC criterion being used to specify the optimal model structure. The results obtained are summarized in Table 2.

At a noise level of 5% the true model structure was identified correctly 98% of the time, with 2% false positive (at least one additional parameter when compared to the true model) models being identified. The MAE and the 2-Norm of the parameter vector $\left\|\hat{\boldsymbol{b}} - \boldsymbol{b}\right\|_2^2$ indicate accurate model estimation (where the MAE was calculated as the mean estimate between the smoothed species concentrations and the estimate of the species concentrations obtained from the MILP). The bias in the parameter values is small and not affected by the false positive results (indicating small parameter values for the additionally identified terms). Similar results are obtained at a noise level of 10% (the correct structure is identified 99%) and 20% (where the correct structure is identified 100% of the time) in both cases with a low training error and accurate parameter estimates. At a noise level of 30%, 94% of the models are estimated with the correct model structure (6% false positive, 4% false negative). The MAE and $\left\|\hat{\boldsymbol{b}} - \boldsymbol{b}\right\|_2^2$ increases as a result of the incorrect model predictions. However, the MILP still performs very well considering the level of noise added to the species measurements. In comparison to Table 1, it is suspected that the performance of the algorithm has improved because of the additional application of the mass balance constraints which reduce the complexity of the optimisation problem by imposing known structural constraints within the network. In

fact, it is conjected that small number of incorrect CRNs identified is attributable to inadequate identification of the smoothing function (19) yielding a poor estimate of the derivative rather than the MILP itself (strengthening our belief that at this crucial step human judgement as to the quality and shape of the smoothed curve are important). Indeed, this might be more critical in situations where there are limited data samples and/or unevenly distributed data.

	MAE	$\left\|\widehat{\boldsymbol{b}} - \boldsymbol{b}\right\|_2^2$	**False Positive (%)**	**False Negative (%)**	**True Model Structure (%)**
~5% Noise	0.0750 ($\pm$ 0.008)	0.0016 ($\pm$ 0.0015)	2	0	98
~10% Noise	0.1071 ($\pm$ 0.0174)	0.0066 ($\pm$ 0.0077)	1	0	99
~20% Noise	0.1938 ($\pm$ 0.0525)	0.0304 ($\pm$ 0.0298)	0	0	100
~30% Noise	0.3007 ($\pm$ 0.1517)	0.2125 ($\pm$ 0.7371)	6	4	94

Table 2. Performance L_0 – norm regularization using an optimal regularization weighting defined via BIC. For each noise level, one hundred randomly generated simulations are used.

7.3 Identification of a transesterification reaction

The formation of biodiesel through the methanolysis of triglycerides is of great interest due to its ability to produce renewable, environmentally friendly fuel from potentially waste materials Ma and Hanna (1999). As a significant amount of research continues in the field, a vast literature is available concerning the kinetics of the process under different conditions and using different raw materials, e.g. Darnoko and Cheryan (2000), Kusdiana and Saka (2001), Singh and Fernando (2007) are just a few examples. Experimental data for this case study was obtained from Almagrbi et al. (2014) who studied the transesterification of sunflower oil at 150^{o}C and 1.0MPa. Reactant species methanol (MeOH) and triglyceride (TG) react to form biodiesel (BD) and glycerol (GL), along with diglyceride (DG) and monoglyceride (MG), which appear as intermediates.

As there are six species, using (6) there are twenty seven possible monomials describing the feasible reactions between species. The elemental matrix, where the rows are ordered as MeOH, TG, DG, BD, MG, GL and the columns are, C, H, O and R (long chain fatty acids) is,

$$\mathbf{A} = \begin{bmatrix} 1 & 3 & 3 & 1 & 3 & 3 \\ 4 & 5 & 6 & 3 & 7 & 8 \\ 1 & 6 & 5 & 2 & 4 & 3 \\ 0 & 3 & 2 & 1 & 1 & 0 \end{bmatrix}^{\mathrm{T}}$$

Using the elemental balance constraints (9) and solving the LP (10) reduces the feasible monomials to a set of ten that can produce feasible (mass balanced reactions),

$$\mathbf{m} = [c_1c_2 \quad c_1c_3 \quad c_1c_5 \quad c_2c_5 \quad c_3c_4 \quad c_3^2 \quad c_4c_6 \quad c_4c_5 \quad c_5c_3 \quad c_6c_2]^{\mathrm{T}}$$

where, $x_1 = \mathrm{MeOH}$, $x_2 = \mathrm{TG}$, $x_3 = \mathrm{DG}$, $x_4 = \mathrm{BD}$, $x_5 = \mathrm{MG}$, $x_6 = \mathrm{GL}$ (and $c_1, \ldots, c_6$ are the species concentrations). Furthermore, the feasible structural interactions within the weighted Laplacian matrix (those that produce balanced chemical reactions) is reduced to ten terms, i.e. ten potential reactions – one reaction corresponding to each row (the Laplacian matrix showing the feasible reactions – the entries indicated with a one – is shown below),

$$\mathbf{L} = \begin{bmatrix} -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix}$$

The species concentration profiles are shown in Fig. 10 along with the smoothed estimate of the output. For this particular example, the best (smooth) approximation (in our judgement) to the measured data was obtained using a smoothing spline. The smoothing spline approximations for each of the species were then used to calculate their respective derivatives. Minimizing (16) subject to the constraints (13) and (15) constraints on the limits of the model constants, $0 \leq k_{pq} \leq 3.5$, the regularization parameter was progressively increased and the BIC criterion calculated - see Fig. 11. The network with the minimum BIC corresponded to regularization values of ($0.007 \leq \lambda \leq 0.43$) and comprised the following three reactions (and their estimated isothermal rate constants),

$$\begin{aligned} &MeOH + TG \rightarrow BD + DG, \quad k_1 = 0.4526\ \mathrm{dm^3.kmol^{-1}.day^{-1}} \\ &MeOH + DG \rightarrow BD + MG, \quad k_2 = 0.3958\ \mathrm{dm^3.kmol^{-1}.day^{-1}} \\ &MeOH + MG \rightarrow BD + GL, \quad k_3 = 0.3523\ \mathrm{dm^3.kmol^{-1}.day^{-1}} \end{aligned} \tag{21}$$

Using L_1-norm regularization (see Fig. 12) the network (21) was also identified as being the optimal structure. The BIC determined weighting in this case was ($0.0009 \leq \lambda \leq 0.017$) and the same numerical values were obtained for the values of the isothermal rate constants (the small regularization weight not contributing to any bias in the parameter values).

A comparison of the response of the process model (21) to the real experimental data may be seen in Fig. 13 (which includes an end-point data sample taken after 6 days of the experimental run which was not used to obtained the estimated derivatives). It may be observed that the simulated data is in agreement with the experimental data. The model estimate of the two reactants (MeOH and TG) as well as the biodiesel (BD) correspond well with the measured concentrations. However, with respect to the intermediate species (DG, MG) and the product (GL) model accuracy could be improved. The difference between the measured and simulated data could be caused by measurement errors, be a facet of the limited number of measurements, or be because the superstructure defined for system identification does not include all possible model terms. The positive aspect of this case study is that it demonstrates the identification of a CRN and kinetic rate constants that agree well with those published in Almagrbi et al. (2014) and the CRN is the accepted structure of biodiesel reactions. However, as with any model development exercise it also points in directions for both further experimental and modelling work in order to improve the accuracy of predictions of DG, MG and GL.

8.0 Discussion and conclusions

The inference of mathematical descriptions of (bio) chemical reaction networks from experimental data is a challenging problem whose solution would have significant commercial and academic impact. It would

allow novel synthetic routes to be characterised, analysed and optimised efficiently using modern simulation tools. This work has focused on the development of a MILP framework for CRN determination. Whilst linear programming and other optimisation techniques have been used to solve many chemical engineering problems for many years, their application in the present field of study has not been documented. The framework was successfully tested using two reaction systems; the Van de Vusse reactions and the methanolysis of triglycerides. The advantages of the approach are **a)** it is almost entirely automated with all parameter and structural interactions are considered simultaneously, **b)** parsimonious solutions are promoted which is advantageous as most (bio) chemical reaction networks are sparsely populated, **c)** unlike typical perturbation methods, it can be applied to chemical reaction systems far from equilibrium, e.g. experiments typically performed in batch reactors, **d)** unlike TFA based methods, it is not necessary to separately determine either the number of independent reactions or reaction stoichiometries, **e)** a priori chemical information in the form of conservation constraints (e.g. the elemental balance) as well as other heuristic constraints (e.g. knowledge of intermediate species) can, when available, be used to both reduce and analyse the structural interactions within the superstructure.

However, there are a number of limitations to the approach in its current form, **a)** the "absolute" optimal solution cannot be obtained if its structure is not covered by the original superstructure; hence in principle, rich and complex superstructures have to be used in order to define a wide search space. Therefore in our future work we wish to establish whether alternative structural constraints may be used to either enhance or eliminate the monomials used to characterize the reactions. It is expected that reaction heat (Q_r) data, e.g. see Wright at al. (1993) may be used to check consistency and develop the structure of the individual species rate expressions (as it is equivalent to the sum of individual heats of reaction multiplied by their reaction rates). This will be particularly important in experimental chemistry where the assumption of mass action kinetics may not apply, **b)** it is assumed that all of the species concentrations are measured whereas in many practical situations this may not be the case. When this is not true, however, it may be possible to employ existing TFA related methods (Bonvin and Rippin, 1990) for estimating concentration data for the unmeasured species, **c)** isothermal reactor operation is assumed. Under controlled laboratory conditions this would normally be true. However, if this were not the case, the method may be extended to non-isothermal identification - through estimation of the Arrenhius coefficients within the ODEs - although this would be at the expense of increased computational complexity defining a mixed integer nonlinear program (MINLP).

In addition, the temporal behaviour of chemical reaction networks is constrained not only by mass flow but also by energy flow. At constant pressure, an isothermal reaction can only proceed in the forward direction if the Gibbs free energy of the reactants consumed is greater than the Gibbs free energy of the products produced e.g. see Beard et al. (2002). Therefore, energy constraints, as well as those based on mass, could be used to enhance the network search process.

Finally, while the emphasis of this paper has been the developed of an overall strategy for the identification of CRNs using MILP the approach adopted for parameter estimation appears to offer a novel regularisation strategy which decouples the parameter estimation and structure identification problems. In our further work, we will investigate and compare this approach with the more traditional approaches used for model regularisation using benchmark data sets from the literature.

9.0 References

Achenie, LEK and Biegler, LT (1990) A superstructure based approach to chemical reactor network synthesis, Comp. Chem. Eng., 14, 23

Agarwal, U. and Singh, UP. (2009) Graph Theory, University Science Press.

Akaike, H. (1974) A new look at the statistical model identification, IEEE Transactions on Automatic Control 19 (6): 716–723

Almagrbi, A., Hatami, T., Glisic S., and Orlovic, A. (2014) Determination of kinetic parameters for complex transesterification reaction by standard optimisation methods, Hem. Ind., 68, 2, 149–159.

Aris R, Mah RHS. (1963) Independence of chemical reactions. Industrial and Engineering Chemistry Fundamentals, 2, 90-94

Bardow and Marquardt 2004, Chemical Engineering Science. 2004; 59(13): 2673–2684.

Beard, DA., Liang, S. and Qian, H. (2002) Energy Balance for Analysis of Complex Metabolic Networks, Biophysical Journal, 83, 1, 79-86.

Bhatt, N, Kerimoglu, N, Amrhein, M, Marquardt, W, Bonvin, D. (2012) Incremental identification of reaction systems - A comparison between rate-based and extent-based approaches, Chem. Eng. Sci. 83, 24-38

Bonvin D, Rippin, DWT. (1990) Target factor analysis for the identification of stoichiometric models. Chem. Eng. Sci., 45(12), 3417-3426.

Brendel, M., Bonvin, D. and Marquardt, W. (2006) Incremental identification of kinetic models for homogeneous reaction systems, Chem. Eng. Sci., 61, 5404-5420

Burnham SC, Searson DP, Willis MJ, Wright AR. (2008) Inference of chemical reaction networks, Chem. Eng. Sci., 63(4), 862-873.

Craciun, G. and Feinberg, M. (2005) Multiple Equilibria in Complex Chemical Reaction Networks: I. The Injectivity Property, SIAM Journal on Applied Mathematics, 65, 1526–1546.

Craciun, G. and Feinberg, M. (2006), Multiple equilibria in complex chemical reaction networks: extensions to entrapped species models, IEE Proceedings - Systems Biology, 153, 179.

Craciun, G. and Feinberg, M. (2010), Multiple Equilibria in Complex Chemical Reaction Networks: Semiopen Mass Action Systems, SIAM Journal on Applied Mathematics, 70, 1859–1877.

Crampin EJ, Schnell S, McSharry PE. (2004) Mathematical and computational techniques to deduce complex biochemical reaction mechanisms. Prog. Biophys. Mol. Biol., 86(1), 77-112.

Darnoko, D. and Cheryan, M. (2000) Kinetics of palm oil transesterification in a batch reactor, J. Am. Oil Chem. Soc., 77, 12, 1263–1267.

Floudas, CA. and Lin, X. (2005) Mixed integer linear programming in process scheduling: modelling algorithms and applications, Annals. of Op. Res., 139, 131 – 162.

Gao, X. (2008) Penalized Methods for High-dimensional Least Absolute Deviations Regression, University of Iowa.

Grossmann, IE. (1985) Mixed integer programming approach for the synthesis of integrated process flowsheets, Comp. Chem. Eng., 9, 5, 463-482

Hadjiconstantinou, E. and Mitra, G. (1994) A linear and discrete programming framework for representing qualitative knowledge, J. of Econ. Dyn. And Control, 18, 273-297.

Hesterberg, T, Choi, NH, Meier, L and Fraley, C. (2008) Least angle and ℓ_1 penalized regression: a review, Statistical surveys, 2, 61-93

Hii, CJK., Wright, AR. and Willis, MJ. (2014) Utilizing a Genetic Algorithm to Elucidate Chemical Reaction Networks: An Experimental Case Study, Int. J. Chem. Eng. Appl., 5, 6, 516–520.

Hoerl, AE. and Kennard, RW. (1970). Ridge regression: Biased estimation for nonorthogonal problems, Technometrics, 12 (1): 55–67

Horn, F. and Jackson, R. (1972) General Mass Action Kinetics, Arch. Ratl. Mech. Anal., 47,81

Hosten, L.H. (1979) A comparative study of short cut procedures for parameter estimation in differential equations, *Computers and Chemical Engineering* 3, 117–126

Kamenski, D.I. and Dimitrov, S.D. (1993) Parameter estimation in differential equations by applications of rational functions, *Computers and Chemical Engineering*, 17 (7), 643–651

Kirchhoff, G. (1847) Ueber die Auflösung der Gleichungen, auf welche man bei der Untersuchung der linearen Verteilung galvanischer Ströme geführt wird, Annalen der Physik und Chemie, 148, 497–508.

Kusdiana, D. and Saka, S. (2001) Kinetics of transesterification in rapeseed oil to biodiesel fuel as treated in supercritical methanol, Fuel, 80, 5, 693–698.

Ma, F. and Hanna, MA. (1999) Biodiesel production: a review, Bioresour. Technol., 70,1,1–15.

Maria, G. (2004) A review of algorithms and trends in kinetic model identification for chemical and biochemical systems, Chem. Biochem. Eng. Q., 18, 3, 195–222.

Marquardt, W. (2005) Model based experimental analysis of kinetic phenomena in multi-phase reactive systems, Chem. Eng. Res. And Des., 83, 561 - 573

Mata-Perez, F. and Perez-Benito, J.F. (1987) The kinetic law for autocatalytic reactions, *Journal of Chemical Education* 64, 11, 925–927.

Mincheva, M (2011) Oscillations in biochemical reaction networks arising from pairs of subnetworks., Bulletin of mathematical biology, 73, 2277–304.

Mincheva, M and Roussel, M. (2007) Graph-theoretic methods for the analysis of chemical and biochemical networks. I. Multistability and oscillations in ordinary differential equation models., Journal of mathematical biology, 55, 61–86.

Moro, LL. and Grossmann, IE. (2013) A Mixed-Integer Model Predictive Control Formulation for Linear Systems, Comp. Chem. Eng., 55, 1-18.

Raman, R. and Grossmann, IE. (1991) Relation between MILP modelling and logical inference for chemical process synthesis, Comp. Chem. Eng., 15, 2, 73-84

Raman, R. and Grossmann, IE. (1992) Integration of Logic and Heuristic Knowledge in the MINLP Optimization for Process Synthesis, Comp. Chem. Eng., 16, 155-171.

Schwarz, GE. (1978) Estimating the dimension of a model, Annals of Statistics, 6 (2): 461–464.

Searson DP, Willis MJ, Horn SJ, Wright AR. (2007) Inference of chemical reaction networks using hybrid s-system models. Chemical Product and Process Modeling, 2 (10).

Searson DP, Willis MJ, Wright A. (2012) Reverse Engineering Chemical Reaction Networks from Time Series Data. *In:* Dehmer, M., Varmuza, K., Bonchev, D, ed. *Statistical Modelling of Molecular Descriptors in QSAR/QSPR*. Weinheim, Germany: Wiley-VCH Verlag GmbH, 327-348.

Shao, J. (1993) Linear model selection by cross-validation, J. Am. Stat. Assoc. 88, 486–494.
Singh, A.K. and Fernando, SD. (2007) Reaction Kinetics of Soybean Oil Transesterification Using Heterogeneous Metal Oxide Catalysts, Chem. Eng. Technol., 30, 12, 1716–1720.
Srividhya J, Crampin EJ, McSharry PE, Schnell S. (2007) Reconstructing biochemical pathways from time course data. Proteomics, 7, 828-838.
Tibshirani, R (1996). 'Regression Shrinkage and Selection via the Lasso', Journal of the Royal Statistical Society, Series B 58 (1), 267–288.
Voit EO, Almeida J. Decoupling dynamical systems for pathway identification from metabolic profiles. Bioinformatics. 2004; 20(11): 1670-1681.
Wang, H., Li, R., and Tsai, C. (2007) Tuning parameter selectors for the smoothly clipped absolute deviation method, Biometrika, 94, 553–568.
Wright, AR., Bramfitt, VJ., Wright, AW. and Zollinger, J. (1993) Kinetic fitting using RC1 data, *RC User Forum*, Zermatt.
Zou, H., Hastie, T., and Tibshirani, R. (2007) On the "degrees of freedom" of the lasso. Ann. Statist. 35.

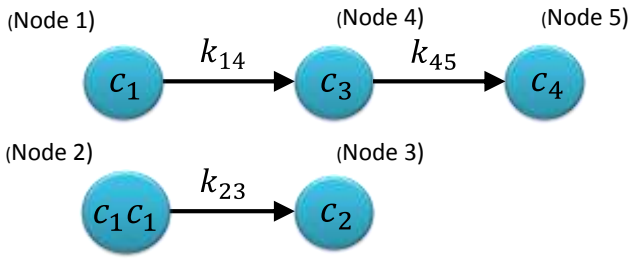


Fig. 1 A weighted digraph representing the Van de Vusse reaction with nodes $\mathbf{m} = [c_1 \quad c_1^2 \quad c_2 \quad c_3 \quad c_4]^{\mathrm{T}}$ edges labelled with isothermal rate constants.

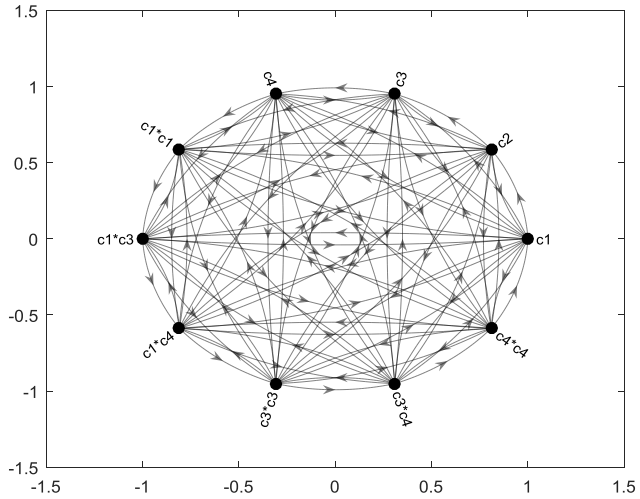


Fig. 2 An unweighted directed graph, representing the reaction superstructure for chemical reactions (1) (all possible reactions between the species - which admits ninety possible chemical reactions). The digraph was generated in MATLAB 2015b using the built in digraph command.

$$\mathbf{LNA} = \begin{bmatrix} -2 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -2 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & -6 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & -2 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & -2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & -2 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & -2 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & -2 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & -2 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 2 \\ 1 \\ 1 \\ 2 \\ 2 \\ 2 \\ 2 \\ 2 \end{bmatrix} = \mathbf{0}$$

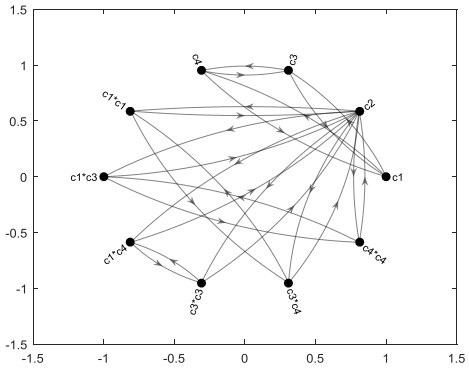


Fig. 3 A directed graph, representing the mass balanced set of feasible reactions and the corresponding unweighted Laplacian matrix. The set of feasible reactions has been reduced to twenty four (a significantly reduced superstructure when compared to Fig. 2).

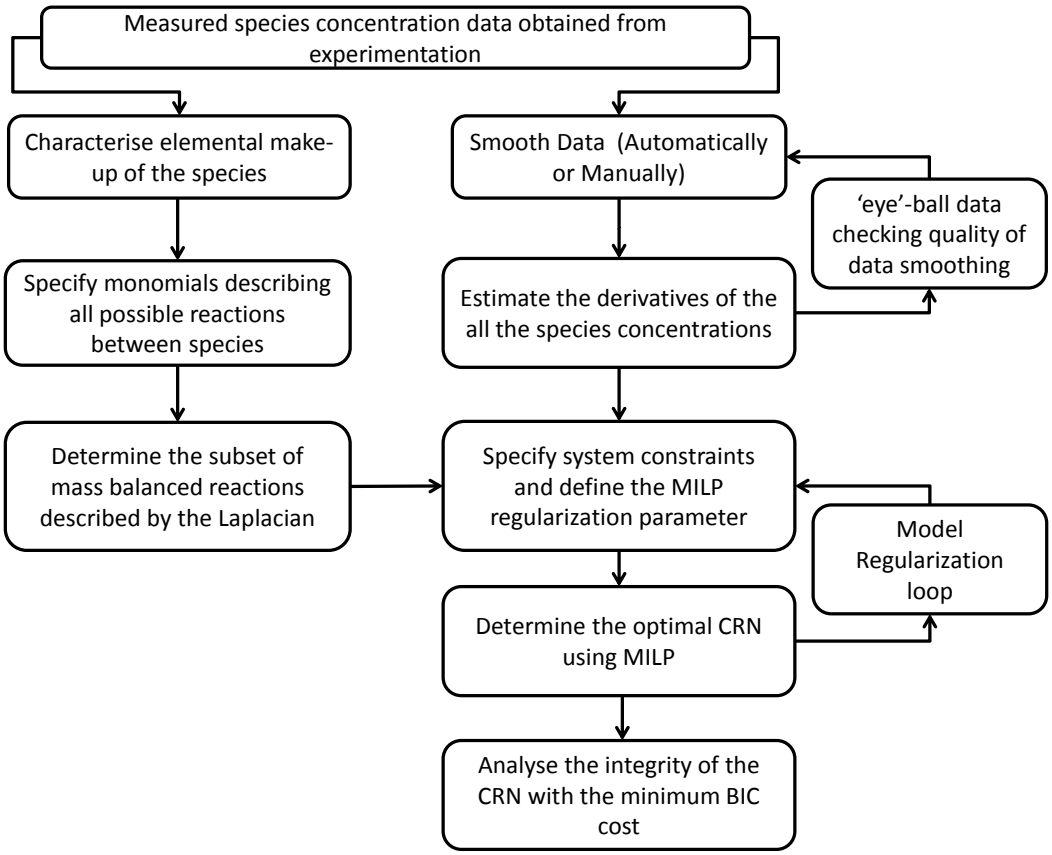
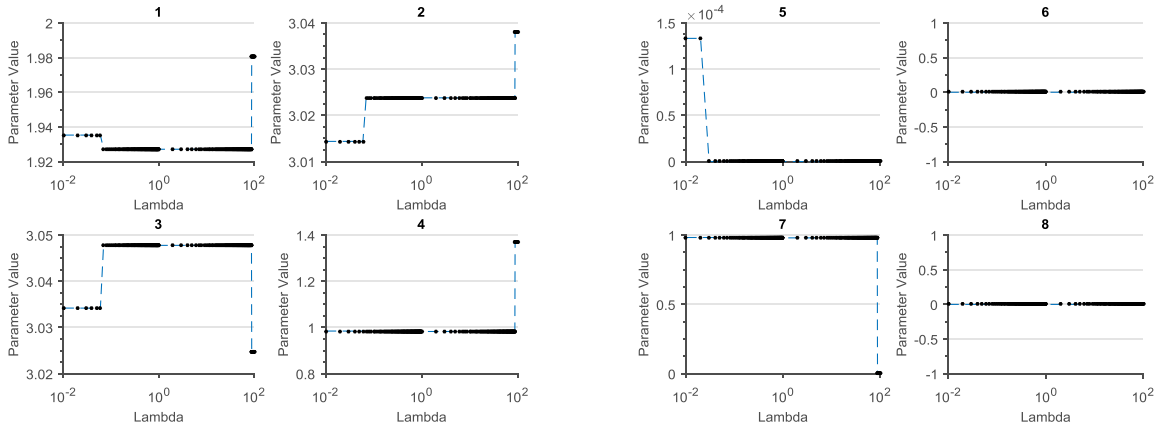


Fig.4 Flowchart of the CRN elucidation strategy, demonstrating the steps involved in the semi-automatic strategy using MILP.



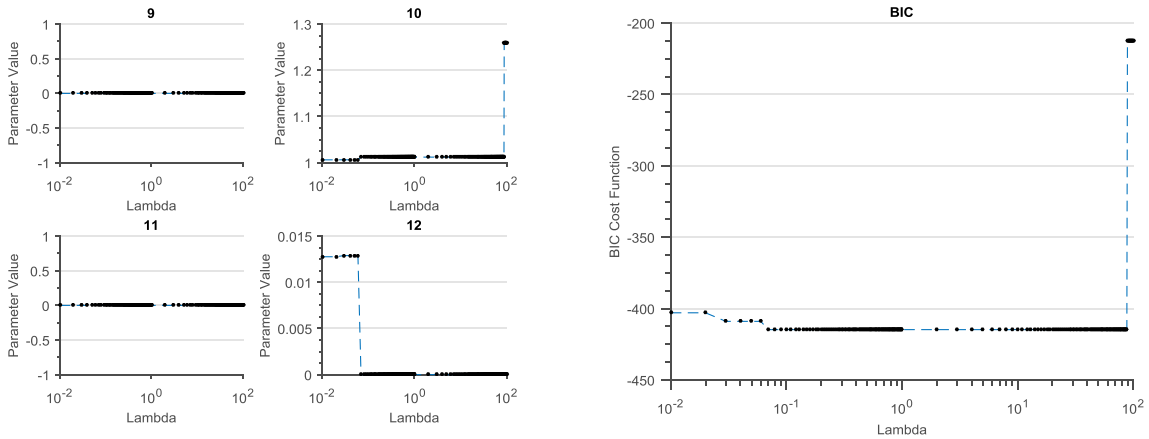


Fig. 5 Regularization path using the L_0 norm weighting for the CRN defined by the Laplacian (19). The numbers (1 – 12) represent the estimated parameters in the Laplacian (ordered down successive rows) as the regularization parameter is increased. The bottom right figure shows the corresponding BIC performance index. The minimum corresponding the regularization parameters in the range, $0.15 \leq \lambda \leq 88$.

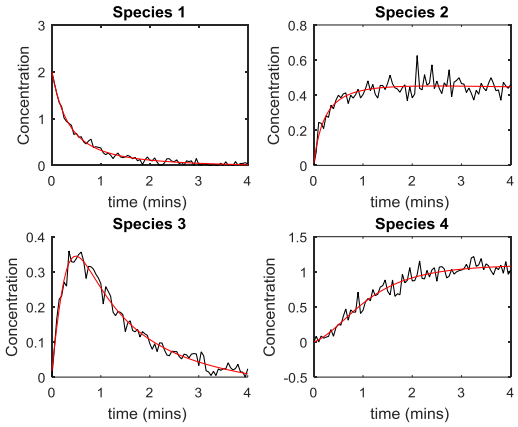


Fig. 6. Species concentrations plotted against time for the Van de Vusse system. The black line corresponds to the noisy (10% white noise) measurement while the red line is the smoothed estimate of the species measurement using (20).

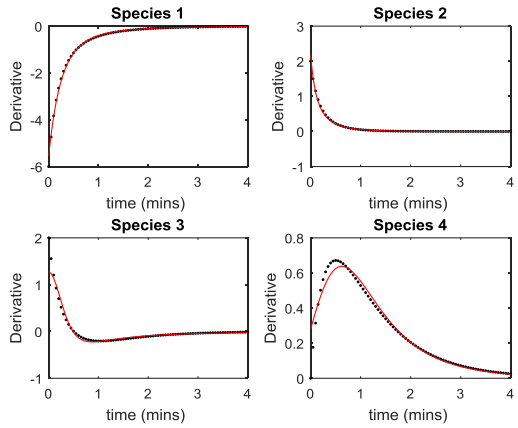


Fig. 7 Derivative of the species concentrations plotted against time for the Van de Vusse system. The true (model) derivative is shown as black dots while the derivative obtained through analytical differentiation of (20) as the red line.

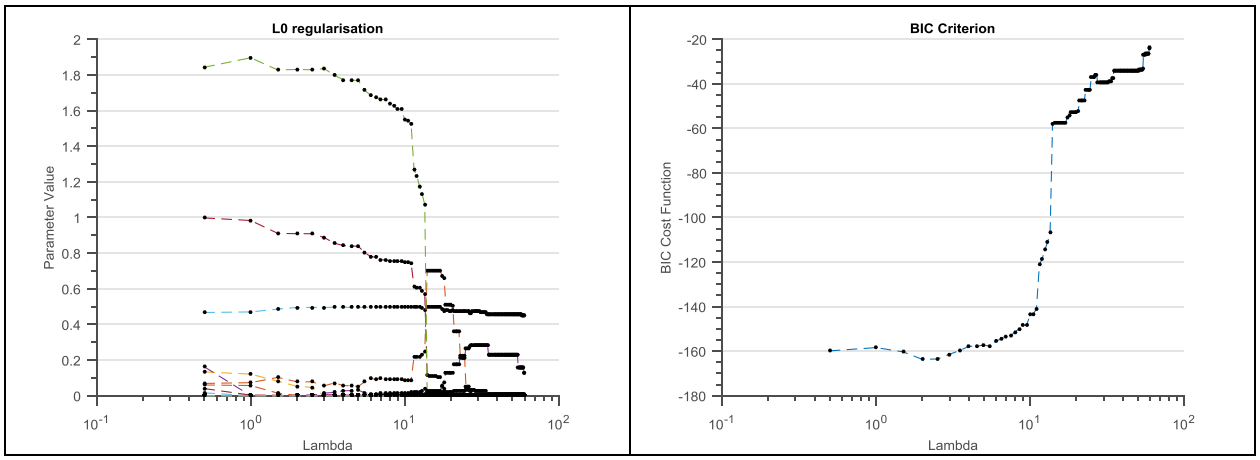


Fig. 8 MILP structure and parameter optimisation using L_1 – norm regularization (application to the Van de Vusse reaction scheme). The data used for identification was the smoothed species measurements shown in Fig. 6 and the corresponding estimate of the derivative shown in Fig. 7. To the left, the transition of the model parameters as the regularization parameter is increased, to the right, the BIC cost function.

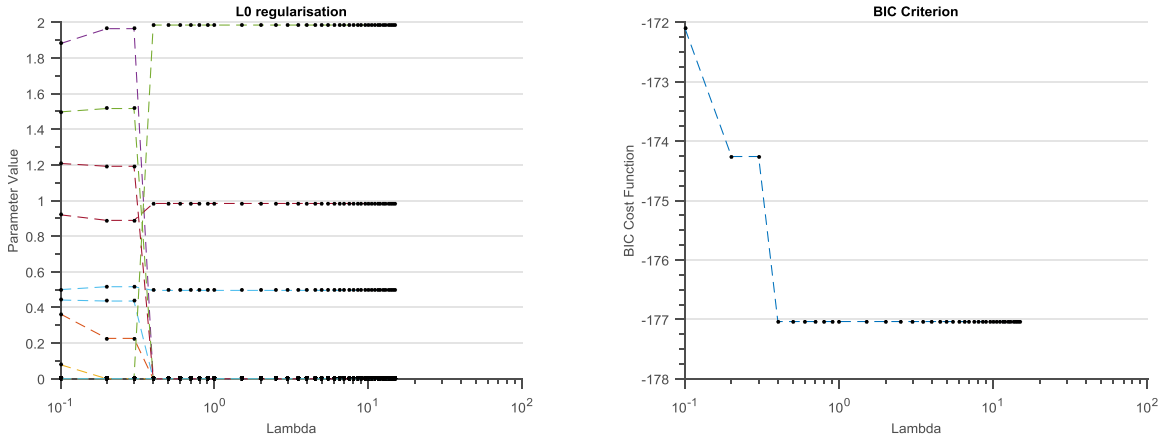


Fig. 9 MILP structure and parameter optimisation using L_0- norm regularization (application to the Van de Vusse reaction scheme). The data used was the smoothed species measurements shown in Fig. 6 and the corresponding estimate of the derivative shown in Fig. 7. The regularization parameter was increased from an initial value of $\lambda = 0.1$ to a final value of 15. To the left, the transition of the model parameters as the regularization parameter is increased, to the right, the BIC cost function.

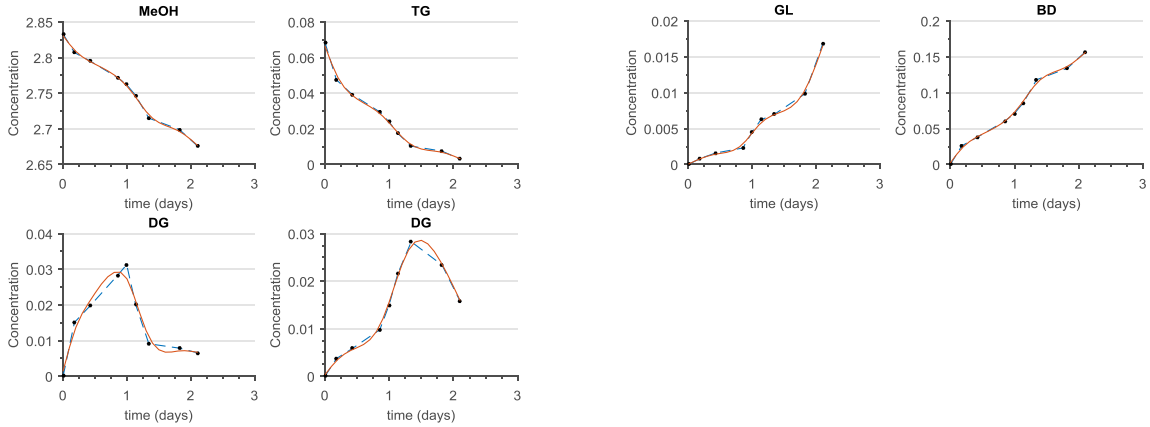


Fig. 10 Experimental data for the transesterification reaction. Blue line (with dots) are the experimental data obtained from Almagrbi et al. (2014). The red line is the smoothed estimate of the experimental data obtained using a smoothing spline.

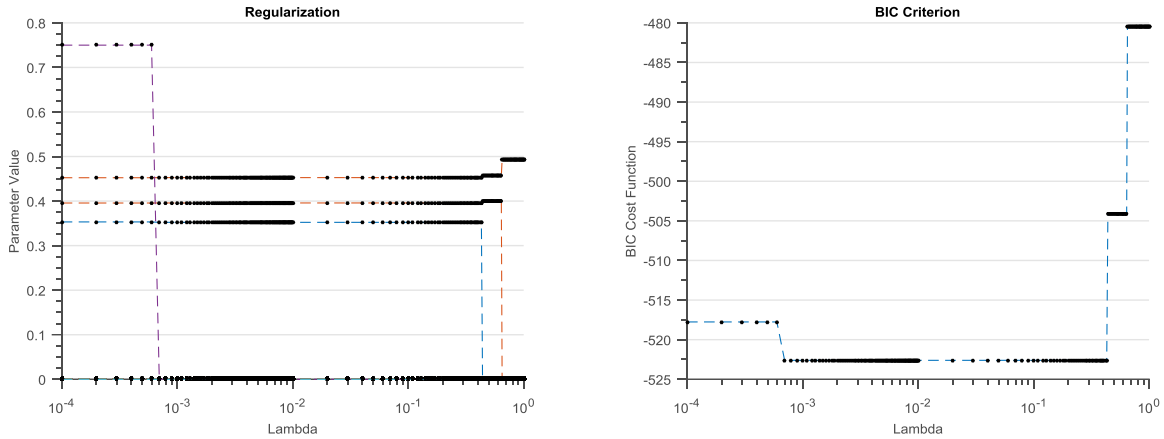


Fig. 11 MILP structure and parameter optimisation using L_0- norm regularization (application to the transesterification reaction data). To the left, the transition of the model parameters as the regularization parameter is increased, to the right, the BIC cost function.

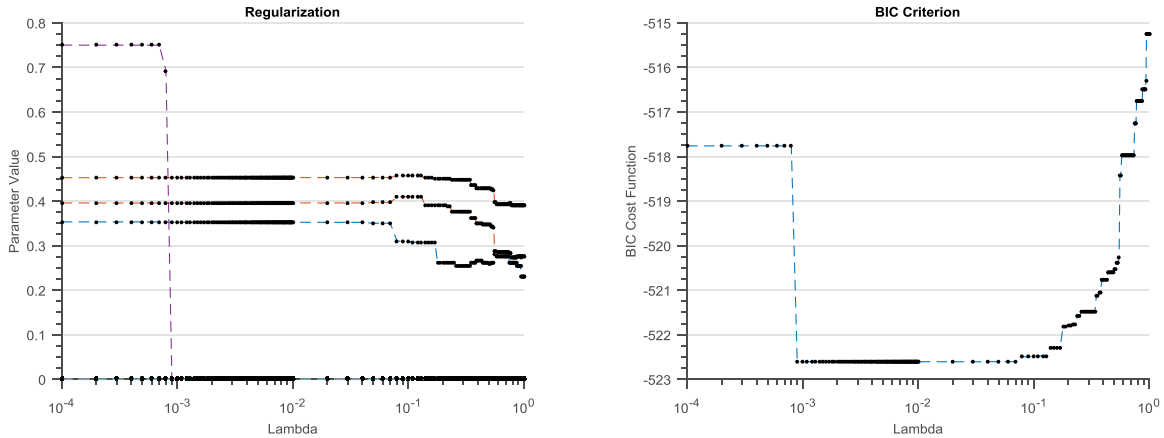


Fig. 12 MILP structure and parameter optimisation using L_1- norm regularization (application to the transesterification reaction data). To the left, the transition of the model parameters as the regularization parameter is increased, to the right, the BIC cost function.

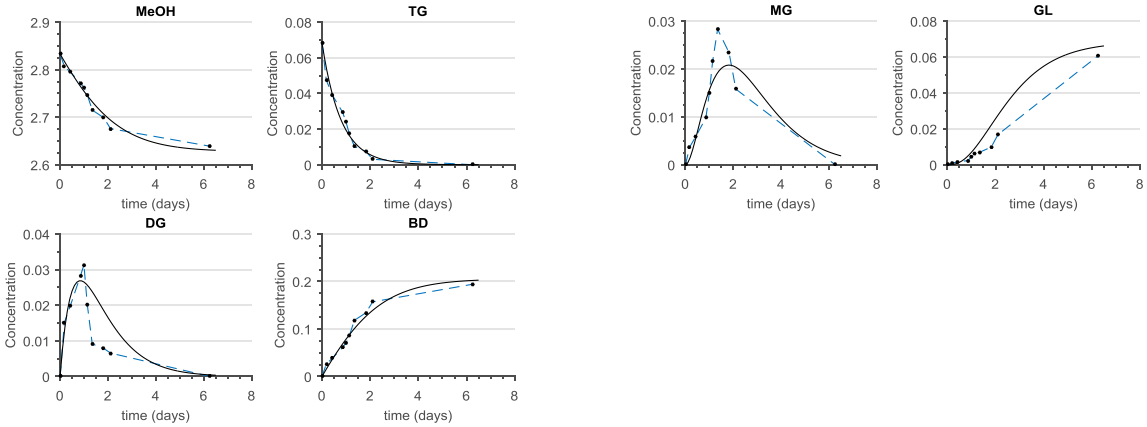


Fig. 13. Comparison of the predicted species concentration profiles and experimental data for the transesterification reaction. Solid black lines are the simulated data, and the blue dashed line with the dots are the experimental data obtained from Almagrbi et al. (2014).